ELSEVIER

CrossMark

Journal of Clinical Epidemiology 75 (2016) 108−114

Journal of Clinical Epidemiology

Empirical evaluation demonstrated importance of validating biomarkers for early detection of cancer in screening settings to limit the number of false-positive findings

Hongda Chen[a], Phillip Knebel[b], Hermann Brenner[a,c,d,*]

[a]*Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, 69120 Heidelberg, Baden-Württemberg, Germany*

[b]*Department of General, Visceral and Transplantation Surgery, University of Heidelberg, Im Neuenheimer Feld 110, 69120 Heidelberg, Baden-Württemberg, Germany*

[c]*Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Im Neuenheimer Feld 460, 69120 Heidelberg, Baden-Württemberg, Germany*

[d]*German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Baden-Württemberg, Germany*

Accepted 25 January 2016; Published online 2 February 2016

Abstract

Objectives: Search for biomarkers for early detection of cancer is a very active area of research, but most studies are done in clinical rather than screening settings. We aimed to empirically evaluate the role of study setting for early detection marker identification and validation.

Study Design and Setting: A panel of 92 candidate cancer protein markers was measured in 35 clinically identified colorectal cancer patients and 35 colorectal cancer patients identified at screening colonoscopy. For each case group, we selected 38 controls without colorectal neoplasms at screening colonoscopy. Single-, two- and three-marker combinations discriminating cases and controls were identified in each setting and subsequently validated in the alternative setting.

Results: In all scenarios, a higher number of predictive biomarkers were initially detected in the clinical setting, but a substantially lower proportion of identified biomarkers could subsequently be confirmed in the screening setting. Confirmation rates were 50.0%, 84.5%, and 74.2% for one-, two-, and three-marker algorithms identified in the screening setting and were 42.9%, 18.6%, and 25.7% for algorithms identified in the clinical setting.

Conclusion: Validation of early detection markers of cancer in a true screening setting is important to limit the number of false-positive findings.

Keywords: Research design; Biomarker identification; Early detection; Validity; Study setting; Methodology

1. Introduction

For most cancers, prognosis strongly varies by stage at diagnosis, and the prospect of cure is much higher when the cancer is detected at an early stage. Search for and validation of biomarkers for early detection of cancer is therefore a very active area of research [1−3]. Ideally, pertinent studies should be conducted in a true screening setting to provide reliable estimates of diagnostic performance in the target population for screening [4−6]. However, this is rarely done in practice for several reasons: First, the prevalence of preclinical cancer overall and of specific cancers in particular in the target population for cancer screening (which typically consists of essentially healthy older adults) is typically very low [7]. Therefore, very large study populations are required to ensure sufficient numbers of cases to estimate sensitivity and other indicators of diagnostic performance with adequate precision. Second, there is often no easy to perform and reliable gold standard

Funding: The BliTz study was partly funded by grants from the German Research Council (DFG, grant no. BR1704/16-1). The work of H.C. was partly supported by China Scholarship Council (CSC). The sponsor had no role in the study design, in the collection, analysis, and interpretation of data.

Conflict of interest: The authors disclose no potential conflicts of interest.

* Corresponding author. Tel./fax: +49-6221-42130.

E-mail address: h.brenner@dkfz-heidelberg.de (H. Brenner).

http://dx.doi.org/10.1016/j.jclinepi.2016.01.022

What is new?

Key findings

- Most studies searching for novel biomarkers for early detection of cancer are conducted in clinical settings, that is, using clinically detected cases. Biomarker levels among such cases may differ from biomarker levels among preclinical cases to be detected by screening for a variety of reasons. Biomarkers identified in such studies may therefore be of questionable use unless they are validated in a true screening setting.
- The authors provide a thorough quantitative illustration of the importance of validation of biomarkers for early detection of cancer in a true screening setting using the example of blood protein markers for early detection of colorectal cancer.

What is the implication and what should change now?

- The interpretation of case–control studies based on patients from clinical settings requires particular caution, due to the potential high proportion of false-positive findings.
- Validation of early detection biomarkers of cancer in a true screening setting is important to limit the number of false-positive findings.

examination (such as colonoscopy for colorectal cancer detection) to which measurements of biomarkers could be compared and that could be applied in such large screening populations.

In practice, it is commonly seen that studies evaluating biomarkers for early detection of cancer therefore recruit a sample of cancer cases in a clinical setting (e.g., newly diagnosed cancer patients admitted to a single or multiple clinics), along with a sample of controls without a known cancer diagnosis [8,9]. Moreover, controls often consist of convenience samples, such as patients from different clinic departments or healthy volunteers, and they often strongly differed from the case groups in many respects, including basic sociodemographic factors [10], such as presence of other diseases or age, which may have important implications for specificity. Under such circumstances, any differences in biomarker levels between cancer patients and controls need to be interpreted with caution because they could simply reflect such differences rather than cancer-related differences. In some studies, matching by key sociodemographic factors, such as sex and age, is used to reduce the risk of such bias, but such matching does not eliminate other sources of differences, such as preceding diagnostic measures that led to the diagnosis of cases or even early treatment. Furthermore, even seemingly perfect matching by factors such as sex and age may sometimes introduce or increase rather than eliminate bias because in a true screening setting, age and sex distribution of those with and without cancer is often not identical [11]. Finally, clinically manifest cases are by definition different from preclinical cancers searched for in a screening setting, and they may differ with respect to a number of factors that favor clinical diagnosis, such as cancer size or stage [10].

It is therefore not surprising that very promising results for the diagnostic performance of cancer early detection markers initially obtained in studies conducted in clinical settings could often not be confirmed in later validations in screening settings. On the other hand, it could be anticipated that good diagnostic performance in screening settings should more often go along with good diagnostic performance in clinical settings. This is because studies conducted in screening populations should primarily identify cancer-related differences (e.g., different expression patterns of tumor-associated biomarkers) between cases and controls which would also be expected to apply to clinical settings. However, evidence on differences in confirmation rates of early detection markers identified in clinical and screening settings from systematic comparative assessment is still sparse. In this study, we provide such an assessment, using the search for blood protein biomarkers for early detection of colorectal cancer as an example.

2. Methods

2.1. Study design

We compared the frequency of initial identification and subsequent validation of protein markers and protein marker combinations indicative of presence of colorectal cancer for the following two scenarios:

1) Use of clinically detected cases in the marker identification set and cases detected in a true screening setting in the validation set.
2) Use of cases detected in a true screening setting in the marker identification set and clinically detected cases in the validation set.

In both scenarios, two sets of participants confirmed to be free of colorectal neoplasms at screening colonoscopy were used as controls.

For this comparison, number and composition of study participants in the clinical setting and the screening setting were kept identical. To achieve statistically robust results, a large number of biomarkers and their combinations were evaluated: 92 single protein markers, $\binom{92}{2}=4,186$ two-marker combinations, and $\binom{92}{3}=125,580$ three-marker combinations.

2.2. Study populations

Subjects from the screening setting were recruited in the context of the BliTz study, which is an ongoing study among participants of screening colonoscopy conducted in cooperation with 20 gastroenterology practices in South-western Germany since November 2005. Detailed information on the BliTz study has been reported elsewhere [12,13]. Briefly, the aim of this study was to evaluate novel tests for early detection of colorectal cancer (CRC). Participants were invited to provide blood and stool samples before the screening colonoscopy. The following exclusion criteria were applied in this analysis to ensure the condition of a true screening setting and to minimize the risk of false-negative results at screening colonoscopy: blood samples taken after screening colonoscopy or blood samples with unknown date of blood withdrawal, history of CRC or inflammatory bowel disease, previous colonoscopy in the last 5 years or unknown colonoscopy history, incomplete colonoscopy, or insufficient bowel preparation (latter two criteria only for controls). From the remaining participants of the BliTz study recruited in 2005–2012 ($N = 4{,}345$), all 35 available cases with newly detected CRC were included in the analysis. For comparison, we included a representative sample of 38 controls free of colorectal neoplasms.

As cases representing the clinical setting, we included 35 clinically detected CRC cases recruited at four hospitals in Southern Germany. Blood samples were withdrawn before any treatment from each participant. Another 38 randomly selected controls free of colorectal neoplasms from the BliTz study were included for comparison. Identical standard operation procedures were used for handling of blood samples in both settings. Plasma samples were centrifuged at 2,123*g* for 10 minutes at 4°C and stored at −80°C until analyses.

Written informed consent was obtained from each participant. All studies were approved by the ethics committees of the medical faculty of the University of Heidelberg and of the respective state medical boards.

2.3. Blood protein biomarkers

Ninety-two predefined human tumor-associated protein biomarkers were measured in 146 samples using Proseek Multiplex Oncology $I^{96\times96}$ (Olink Bioscience, Uppsala, Sweden) (full marker list is provided in Supplementary Table S1 at www.jclinepi.com). All laboratory operations were conducted according to the Proseek Multiplex Oncology $I^{96\times96}$ User Manual in the TATAA Biocenter (Göteborg, Sweden) [14]. In short, the Proseek reagents are based on the Proximity Extension Assay technology [15], where 92 oligonucleotide labeled antibody probe pairs are allowed to bind to their respective target present in the sample. A polymerase chain reaction (PCR) reporter sequence is formed by a proximity dependent DNA polymerization event and is subsequently detected and quantified using real-time PCR. All information regarding the study population was blind to the laboratory operators. Two independent analyses were conducted: 35 CRC samples recruited in the clinical setting and 38 controls were measured in a first analysis, and the 35 CRC samples and 38 controls recruited in the screening setting were measured in a second analysis.

2.4. Statistical analysis

Age, sex, and tumor stage distribution (CRC cases only) are reported by descriptive statistics. As identification of biomarkers showing different distributions of biomarker values is typically the first step in the early phase of biomarker research for cancer early detection, we used Wilcoxon rank sum test to identify single biomarkers showing statistically significant differences in plasma levels of CRC cases and controls. Additionally, receiver operating characteristics (ROC) analysis was conducted, and the areas under the ROC curves are reported. Two- and three-marker combinations were evaluated by multiple logistic regression models, and likelihood ratio tests comparing the full model (including the two markers or three markers and intercept) and the null model (including the intercept only) were used to test for statistical significance. Apart from unadjusted *P*-values, *P*-values adjusted for multiple testing using the Benjamini–Hochberg method are also reported [16]. Furthermore, to eliminate potential confounding by age and sex, results from age and sex adjusted logistic regression models are also reported.

All statistical analyses were performed with the statistical software R version 3.1.1 [17]. All tests were two sided, and *P*-values less than 0.05 or false discovery rates (FDRs) less than 5% were considered to indicate statistical significance.

3. Results

Table 1 presents main characteristics of colorectal cancer cases recruited in the clinical setting and the screening setting and their controls. Sex, age, and stage distribution were roughly comparable between both samples. There were more men than women in both CRC patient groups. The control groups included more women, and the proportions in the both control groups were the same (57.9%). Patients with CRC were slightly older than controls in both study samples. The differential sex and age distributions of cases and controls reflect the sex distributions encountered in colonoscopy screening in Germany [10].

Table 2 shows the results of single biomarkers that were found to show statistically significant differences in the clinical setting, with replication in the screening setting. Overall, 25 biomarkers were identified to show statistically significant differences of plasma levels between CRC cases recruited in the clinical setting and controls, but only six of them (6/25 = 24%) were confirmed in the screening

Table 1. Characteristics of colorectal cancer cases recruited in the clinical setting and the screening setting and their controls

Sample characteristics	Clinical setting		Screening setting	
	Colorectal cancer	Controls	Colorectal cancer	Controls
N	35	38	35	38
Men [*N* (%)]	22 [62.9]	16 [42.1]	25 [71.4]	16 [42.1]
Age [mean (SD) years]	68.0 [8.5]	60.7 [6.0]	66.9 [6.5]	62.7 [7.1]
TNM tumor stage [*N* (%)]				
Early stage (stage I/II)	19 [54.3]	–	17 [48.5]	–
Advanced stage (stage III/IV)	16 [45.7]	–	18 [51.4]	–

Abbreviation: SD, standard deviation; TNM, tumor-node-metastasis.

setting. When using 5% FDR as the cutoff level for multiple testing, of seven single biomarkers still showing statistically significant results in the clinical setting, only 3 biomarkers (3/7 = 43%) were successfully replicated in the screening setting samples.

Table 3 lists the results of single biomarkers that were found to show statistically significant differences in the screening setting, with replication in the clinical setting. Overall, of 15 single markers showing statistically significant results in the screening setting, six biomarkers (6/15 = 40%) were successfully replicated in the clinical setting. When using 5% FDR as the cutoff level for multiple testing, of four biomarkers showing statistically significant results, two biomarkers (50%) were confirmed in the clinical setting.

Detailed comparisons of results on two- and three-marker combinations are shown in Table 4. These comparisons confirm, for the much larger numbers of marker combinations, the above described pattern: higher number of initially identified markers but lower subsequent confirmation proportions when identification is done in the clinical setting compared to marker identification in the screening setting. The confirmation rates (results after adjustment for multiple testing) for two- and three-marker combinations identified in the clinical setting were 18.6% (221/1,188) and 25.7% (12,927/50,291), respectively. In the screening setting, smaller numbers of significant marker combinations were initially identified, but much higher confirmation rates were achieved. The confirmation rates were 84.5% (239/283) and 74.2% (11,653/15,703) for

Table 2. Protein markers with significant differences in clinical setting ($P < 0.05$) and their replication in the screening setting

Marker	Identification in the clinical setting			Replication in the screening setting		
	AUC	Unadjusted *P*-value	Adjusted *P*-value[a]	AUC	Unadjusted *P*-value	Adjusted *P*-value[a]
Adrenomedullin	0.81	**<0.001**	**<0.001**	0.61	**0.099**	0.118
Amphiregulin	0.81	**<0.001**	**<0.001**	0.74	**<0.001**	**0.002**
GDF-15	0.81	**<0.001**	**<0.001**	0.72	**0.001**	**0.004**
Cathepsin-D	0.75	**<0.001**	**0.005**	0.62	0.073	0.118
PlGF	0.74	**<0.001**	**0.005**	0.61	0.101	0.118
EGFR	0.73	**<0.001**	**0.007**	0.59	0.196	0.196
IL-6	0.71	**0.002**	**0.027**	0.70	**0.004**	**0.009**
HGF	0.69	**0.005**	0.056	0.53	0.665	—
Follistatin	0.68	**0.008**	0.076	0.60	0.160	—
TNF-RI	0.68	**0.008**	0.076	0.56	0.376	—
Osteoprotegerin	0.68	**0.009**	0.076	0.56	0.382	—
ErbB3-Her3	0.66	**0.016**	0.110	0.59	0.192	—
HE4	0.66	**0.016**	0.110	0.58	0.220	—
CSF-1	0.66	**0.017**	0.110	0.50	1.000	—
MIC-A	0.66	**0.019**	0.116	0.51	0.925	—
PSA	0.65	**0.020**	0.116	0.67	**0.013**	—
VEGF-A	0.65	**0.023**	0.123	0.51	0.926	—
VEGFR-2	0.65	**0.024**	0.123	0.64	**0.043**	—
TNF-R2	0.65	**0.028**	0.135	0.65	**0.032**	—
CXCL13	0.65	**0.030**	0.139	0.59	0.167	—
Flt3L	0.65	**0.032**	0.140	0.59	0.203	—
CXCL11	0.64	**0.038**	0.158	0.60	0.138	—
Midkine	0.64	**0.042**	0.162	0.55	0.479	—
Prostasin	0.64	**0.042**	0.162	0.56	0.382	—
REG-4	0.64	**0.045**	0.164	0.55	0.432	—

Abbreviation: AUC, area under the receiver operating characteristic curve.

Bold indicates statistically significant.

[a] The *P*-value was adjusted for multiple testing using Benjamini–Hochberg method. Statistical significance was defined as false discovery rate <5%.

Table 3. Protein markers with significant differences in screening setting ($P < 0.05$) and their replication in the clinical setting

	Identification in the screening setting			Replication in the clinical setting		
Marker	**AUC**	**Unadjusted *P*-value**	**Adjusted *P*-value[a]**	**AUC**	**Unadjusted *P*-value**	**Adjusted *P*-value[a]**
Amphiregulin	0.74	**<0.001**	**0.021**	0.81	**<0.001**	**<0.001**
CXCL9	0.73	**0.001**	**0.021**	0.62	0.082	0.054
CEA	0.73	**0.001**	**0.021**	0.63	0.051	0.068
GDF-15	0.72	**0.001**	**0.028**	0.81	**<0.001**	**<0.001**
IL-6	0.70	**0.004**	0.062	0.71	**0.002**	—
CXCL10	0.69	**0.004**	0.062	0.57	0.314	—
PSA	0.67	**0.013**	0.161	0.65	**0.020**	—
IFN-gamma	0.67	**0.014**	0.161	0.56	0.413	—
ErbB4-Her4	0.66	**0.017**	0.171	0.56	0.359	—
TNFRSF4	0.65	**0.031**	0.254	0.60	0.163	—
TNF-R2	0.65	**0.032**	0.254	0.65	**0.028**	—
CA-125	0.64	**0.033**	0.254	0.61	0.098	—
VEGFR-2	0.64	**0.043**	0.285	0.65	**0.024**	—
E-selectin	0.64	**0.043**	0.285	0.50	0.987	—
TNF-alpha	0.63	**0.048**	0.295	0.58	0.213	—

Abbreviation: AUC, area under the receiver operating characteristic curve.
Bold indicates statistically significant.
[a] The *P*-value was adjusted for multiple testing using the Benjamini–Hochberg method. Statistical significance was defined as false discovery rate <5%.

two- and three-marker combinations, respectively. When adjusting for age and sex, numbers of identified and confirmed markers and marker combinations were reduced. Nevertheless, higher confirmation rates were still observed for marker combinations derived from the screening setting than for those derived from the clinical setting.

4. Discussion

Our empirical example demonstrates that using samples from the clinical setting for identification of cancer early detection biomarkers might result in a large amount of false-positive findings which cannot be reproduced in a true screening setting. In our example, the confirmation rates (results adjusting for multiple testing) of single-, two-, and three-marker combinations identified initially in the clinical setting were only 42.9%, 18.6% and 25.7%, respectively, if subsequently validated in the screening setting. These confirmation rates were much lower than the confirmation rates of markers and marker combinations identified in a true screening setting. Our results underline the necessity of validating promising biomarker findings in prospective samples from screening settings to limit the number of false-positive findings.

Biomarker research on early detection of cancer is blossoming. However, most present studies are still in the early phase aiming for marker discovery [18]. Strong claims of novel biomarkers for early diagnosis are commonly reported, but often show weak or no reproducibility [19,20]. Part of this phenomenon might be due to the widely used case–control study design in the biomarker discovery phase [8,21]. In such a study design, cancer cases are typically symptomatic patients recruited from hospitals and corresponding controls are often convenience samples, such as patients with different diseases recruited from the same hospital, blood donors, or other healthy volunteers.

There are numerous reasons why apparently most promising biomarkers identified in such settings may often not be confirmed in subsequent validation in true screening settings. First, clinically confirmed diagnoses of cancer recruited in clinical settings may differ from preclinical cases searched for in screening settings with respect to a

Table 4. The comparison of significant single biomarkers, two-, and three-marker combinations between the two settings

		Single marker (*N* = 92)		Two-marker combination (*N* = 4,186)		Three-marker combination (*N* = 125,580)	
Approaches (training set → validation set)	**Adjustment**	**Number**	**%[a]**	**Number**	**%[a]**	**Number**	**%[a]**
Clinical setting → screening setting	No	25 → 6	24.0	1,834 → 633	34.5	66,332 → 26,507	40.0
Screening setting → clinical setting	No	15 → 6	40.0	957 → 633	66.1	38,392 → 26,507	69.0
Clinical setting → screening setting	Multiple testing[b]	7 → 3	42.9	1,188 → 221	18.6	50,291 → 12,927	25.7
Screening setting → clinical setting	Multiple testing[b]	4 → 2	50.0	283 → 239	84.5	15,703 → 11,653	74.2
Clinical setting → screening setting	Age and sex	10 → 2	20.0	881 → 179	20.3	33,054 → 7,927	24.0
Screening setting → clinical setting	Age and sex	7 → 2	28.6	584 → 179	30.7	21,481 → 7,927	36.9

[a] Percentage of markers that could be replicated in the validation setting.
[b] The *P*-value was adjusted for multiple testing using the Benjamini–Hochberg method. Statistical significance was defined as false discovery rate <5%.

large number of relevant factors [10], such as sociodemographic factors (e.g., age, sex), tumor characteristics (e.g., stage at diagnosis, location, grade, and so forth), and clinical factors (e.g., diagnosis and treatment-related factors). Second, convenience samples may likewise often substantially differ from controls recruited in a true screening setting with respect to sociodemographic factors or factors related to their specific diseases (when controls with another disease are used). Third, sample handling and preanalytical sample processing may often substantially differ between samples collected in clinical and screening settings [22,23]. Fourth, apparent diagnostic performance of biomarkers identified in a specific data set may generally be overoptimistic and often not be replicated unless adequate measures are taken [24], such as correction for multiple testing (especially for studies testing a large number of biomarkers simultaneously in a limited number of samples) [25,26], internal validation (i.e., bootstrap and cross-validation) [27], or external validation in an independent data set [27,28].

In our analyses, we already minimized the potential for bias by most of these sources. In particular, we used identical standard operation procedures for obtaining and processing blood samples in both the clinical and the screening setting. Moreover, we used comparable control groups from a true screening setting for both groups of cases. Although these control groups differed from the case groups with respect to age and sex distribution to some extent, this difference reflects the true difference between cases and controls to be expected in a true screening setting and does not introduce bias when judging the performance of early detection markers in such a setting [11]. Despite minimizing the aforementioned sources of bias, initial discovery of biomarkers using cases recruited in the clinical setting still led to substantially larger numbers of false-positive findings that could not be confirmed in subsequent validation. Mechanisms contributing to this pattern might include, for example, differences in tumor characteristics other than stage (whose distribution was roughly similar in our study between cases recruited in the clinical and the screening setting, respectively) or influences of the preceding diagnostic process (e.g., CRC patients recruited in the clinical setting typically have had a colonoscopy at which the cancer was diagnosed days to weeks before recruitment). Regardless of the contributing mechanisms, initial discovery of markers in the clinical setting resulted in substantially higher initial false-positive rates than initial discovery in the screening setting. This pattern would be expected to be even more pronounced in studies taking less care to overcome other aforementioned potential sources of bias.

Despite the expected higher initial false positivity rate, it may often be reasonable to start biomarker discovery using patients recruited in clinical settings [18,19]. The main reason is that identification of sufficiently large numbers of cases in true screening settings, if feasible at all, typically requires recruitment of very large screening cohorts, given the low prevalence of most preclinical cancers. The effort for setting up such a cohort or using the precious samples from existing screening cohorts may not be worthwhile when biomarkers turn out to show poor diagnostic performance even when using clinically detected cases. Discovery of candidate biomarkers in a clinical setting, which typically is much less time consuming and costly, may therefore often be a reasonable first step. Our analyses underline, however, the importance of subsequent validation of findings in a true screening setting to limit the number of false-positive results. Based on a novel analysis, we quantitatively report the impact of study settings on the blood biomarker identification for early detection of cancer. There are specific strengths and limitations that deserve careful consideration when interpreting our results. Strengths include that a large number of biomarkers were simultaneously measured in samples of equal size from both a clinical setting and a screening setting, which enabled a fair comparison of biomarker identification in both settings. We did not restrain our analysis on single markers only, but also evaluated all two- and three-marker combinations, which yielded a very broad empirical basis for our analyses. In addition, the samples from the screening setting were subgroups of participants recruited in a very large cohort of screening colonoscopy, which is an ideal target population for evaluating the diagnostic performance of novel biomarkers for early detection of CRC.

The limitations include the relatively small sample size of CRC cases in the screening setting although a very large number of participants were recruited, which reflects the very low prevalence of CRC in a screening population. The same number of cases was selected for the clinical setting and for the controls. Sample sizes of this order are common in studies evaluating cancer early detection markers [8,21]. There were slight differences regarding age and sex distribution between the CRC cases used in the two settings. As age and sex are known risk factors of CRC [29], they could also have affected the confirmation rates calculated in our analysis. Moreover, only protein biomarkers were evaluated in our study. Further studies using other technology or evaluating other biomarkers would be desirable to supplement our findings. Our analyses only evaluated CRC as disease outcome. We would anticipate, however, that similar findings would also apply for other cancers. Although our study demonstrated major difference in confirmation rates according to study settings, the overall diagnostic performance of the biomarkers assessed in this study was lower than would be desirable for a CRC screening test. However, similar difference in confirmation rates according to the study setting might also occur for tests with higher diagnostic performance.

In conclusion, study designs may strongly affect the validity of studies on biomarkers for early detection of cancer. The interpretation of case–control studies based on patients from clinical settings requires particular caution, due to the potential high proportion of false-positive findings.

Validation of biomarkers for early detection of cancer in screening settings is important to limit the number of false-positive findings.

Acknowledgments

The authors gratefully acknowledge the excellent cooperation of gastroenterology practices and clinics in patient recruitment and of Labor Limbach in sample collection. The authors gratefully acknowledge Dr. Katja Butterbach and Ulrike Schlesselmann for their excellent work in laboratory preparation of blood samples. The authors also gratefully acknowledge Isabel Lerch, Susanne Köhler, Utz Benscheid, Jason Hochhaus, and Maria Kuschel for their contribution in data collection, monitoring, and documentation.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.01.022.

References

[1] Coghlin C, Murray GI. Biomarkers of colorectal cancer: recent advances and future challenges. Proteomics Clin Appl 2015;9:64−71.
[2] Hasan N, Kumar R, Kavuru MS. Lung cancer screening beyond low-dose computed tomography: the role of novel biomarkers. Lung 2014;192:639−48.
[3] Jenkinson C, Earl J, Ghaneh P, Halloran C, Carrato A, Greenhalf W, et al. Biomarkers for early diagnosis of pancreatic cancer. Expert Rev Gastroenterol Hepatol 2014;9:1−11.
[4] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926−30.
[5] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003;56:1118−28.
[6] Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45:1143−54.
[7] Gatta G, Mallone S, van der Zwan JM, Trama A, Siesling S, Capocaccia R. Cancer prevalence estimates in Europe at the beginning of 2000. Ann Oncol 2013;24:1660−6.
[8] Chen H, Werner S, Tao S, Zornig I, Brenner H. Blood autoantibodies against tumor-associated antigens as biomarkers in early detection of colorectal cancer. Cancer Lett 2014;346:178−87.
[9] Zaenker P, Ziman MR. Serologic autoantibodies as diagnostic cancer biomarkers−a review. Cancer Epidemiol Biomarkers Prev 2013;22:2161−81.
[10] Tao S, Hundt S, Haug U, Brenner H. Sensitivity estimates of blood-based tests for colorectal cancer detection: impact of overrepresentation of advanced stage disease. Am J Gastroenterol 2011;106:242−53.
[11] Brenner H, Altenhofen L, Tao S. Matching of controls may lead to biased estimates of specificity in the evaluation of cancer screening tests. J Clin Epidemiol 2013;66:202−8.
[12] Brenner H, Tao S, Haug U. Low-dose aspirin use and performance of immunochemical fecal occult blood tests. JAMA 2010;304:2513−20.
[13] Hundt S, Haug U, Brenner H. Comparative evaluation of immunochemical fecal occult blood tests for colorectal adenoma detection. Ann Intern Med 2009;150:162−9.
[14] Schneider MR, Hoeflich A, Fischer JR, Wolf E, Sordat B, Lahm H. Interleukin-6 stimulates clonogenic growth of primary and metastatic human colon carcinoma cells. Cancer Lett 2000;151:31−8.
[15] Lundberg M, Eriksson A, Tran B, Assarsson E, Fredriksson S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. Nucleic Acids Res 2011;39:e102.
[16] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological) 1995;57:289−300.
[17] Bolocan A, Ion D, Ciocan DN, Paduraru DN. Prognostic and predictive factors in colorectal cancer. Chirurgia (Bucur) 2012;107:555−63.
[18] Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054−61.
[19] Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. J Clin Epidemiol 2007;60:1205−19.
[20] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
[21] Werner S, Chen H, Tao S, Brenner H. Systematic review: serum autoantibodies in the early detection of gastric cancer. Int J Cancer 2015;136:2243−52.
[22] Behrens T, Bonberg N, Casjens S, Pesch B, Bruning T. A practical guide to epidemiological practice and standards in the identification and validation of diagnostic markers using a bladder cancer example. Biochim Biophys Acta 2014;1844:145−55.
[23] Ransohoff DF, Gourlay ML. Sources of bias in specimens for research about molecular markers for cancer. J Clin Oncol 2010;28:698−704.
[24] Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics 2004;5(6):709−19.
[25] Bender R, Lange S. Adjusting for multiple testing−when and how? J Clin Epidemiol 2001;54:343−9.
[26] Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol 2014;67:850−7.
[27] Taylor JM, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. Clin Cancer Res 2008;14:5977−83.
[28] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25−34.
[29] Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet 2014;383:1490−502.